(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
17 April 2003 (17.04.2003)

PCT

(10) International Publication Number
WO 03/032123 A2

(51) International Patent Classification[7]: G06F

(21) International Application Number: PCT/US02/32234

(22) International Filing Date: 9 October 2002 (09.10.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/975,769 11 October 2001 (11.10.2001) US

(63) Related by continuation (CON) or continuation-in-part (CIP) to earlier application:
US 09/975,769 (CON)
Filed on 11 October 2001 (11.10.2001)

(71) Applicant *(for all designated States except US)*: **PROFITLOGIC, INC.** [US/US]; Eleven Cambridge Center, Cambridge, MA 02142-1406 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **KUMAR, Mahesh** [IN/US]; 305 Memorial Drive, Apt. #404C, Cambridge, MA 02139 (US). **GAIDAREV, Peter** [RU/US]; 20 Summer Street #603N, Malden, MA 02148 (US). **WOO, Jonathan, W.** [US/US]; 101 Monmouth Street, #202, Brookline, MA 02446 (US).

(74) Agent: **POWSNER, David, J.**; Nutter McClennen & Fish LLP, 155 Seaport Boulevard, World Trade Center West, Boston, MA 02110-2604 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii)) for the following designations AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZM, ZW, ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG)*

— *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii)) for the following designations AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZM, ZW, ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG)*

Published:

— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: CLUSTERING

(57) Abstract: A set of data is received containing values associated with respective data points, the values associated with each of the data points being characterized by a distribution. The values for each of the data points are expressed in a form that includes information about a distribution of the values for each of the data points. The distribution information is used in clustering the set of data with at least one other set of data containing values associated with data points.

WO 03/032123 A2

CLUSTERING

BACKGROUND

This invention relates to clustering, for example, of seasonalities to better forecast demand for items of commerce.

Clustering means grouping objects, each of which is defined by values of attributes, so that similar objects belong to the same cluster and dissimilar objects belong to different clusters. Clustering has applications in many fields, including medicine, astronomy, marketing, and finance.

Clustering is done on the assumption that attribute values representing each object to be clustered are known deterministically with no errors. Yet, often, the values representing an object to be clustered are not available. Sometimes statistical methods are used to get estimated or average values for a given object.

SUMMARY

In general, in one aspect, the invention features a method that includes (a) receiving a set of data containing values associated with respective data points, the values associated with each of the data points being characterized by a distribution, (b) expressing the values for each of the data points in a form that includes information about a distribution of the values for each of the data points, and (c) using the distribution information in clustering the set of data with at least one other set of data containing values associated with data points.

Implementations of the invention may include one or more of the following features. The respective data points are related in a time-sequence. The data points relate to a seasonality of at least one item of commerce. Each of the sets of data relates to seasonalities of items of commerce. The items of commerce comprise retail products, the data points relate to times during a season, and the values associated with each of the data points correspond to respective ones of the retail products. The method also includes determining statistical measures of the variability of the values with respect to the data point. The data is expressed in a form that includes a mean of the values associated with a data point and a statistical measure of the distribution with respect to the mean. The statistical measure comprises a standard deviation. The clustering of data includes measuring a distance between pairs of the sets of data. The distance is measured based on the means and variances at the data points. The distribution of the values is Gaussian. The clustering of data includes merging the data sets belonging to a cluster using a weighted average. The method includes merging the seasonalities of the data sets belong to a cluster.

In general, in another aspect, the invention features a machine-accessible medium that when accessed results in a machine performing operations that include: (a) receiving a set of data containing values associated with respective data points, the values associated with each of the data points being characterized by a distribution, (b) expressing the values for each of the data points in a form that includes information about a distribution of the values for each of the data points, and (c) using the distribution

information in clustering the set of data with at least one other set of data containing values associated with data points.

In general, in another aspect, the invention features a method that includes (a) receiving sets of data, each of the sets containing values associated with respect data points, the values associated with each of the data points being characterized by a distribution, (b) evaluating a distance function that characterizing the similarity or dissimilarity of at least two of the sets of data, the distance function including a factor based on the distributions of the values in the sets, and (c) using the evaluation of the distance function as a basis for clustering of the sets of data.

In general, in another aspect, the invention features a method that includes (a) receiving data that represents seasonality time-series for each of a set of retail items, the data also representing error information associated with data values for each of a series of time points for each of the items, and (b) forming composite seasonality time-series based on respective clusters of the retail item seasonality time-series, the composites formed based in part on the error information.

Other advantages and features will become apparent from the following description and from the claims.

DESCRIPTION

(Figure 1 shows product life cycle curves.

Figure 2 shows a seasonality curve and two sales curves.

Figure 3 shows a concatenation of PLC curves and an average curve.

Figure 4 shows three seasonality curves.

Figure 5 illustrates the merger of time-series.

Figure 6 illustrates the merger of time-series.

Figures 7, 8, 9, 10 and 11 show seasonality curves.

Figure 12 is a flow chart.)

Introduction

The clustering of objects can be improved by incorporating information about errors, which are a measure of level of confidence, in data that characterize the objects to be clustered. Instead of relying on simple average values for attributes of the objects, measures of error such as standard deviation or an estimate of an entire distribution of the data for a value can be used to improve the quality of the clustering. A new distance function that is based on the distribution of errors in data (and may be viewed as a generalization of the classical Euclidian distance function) can be used in clustering.

Although the following discussion focuses on the utility of the new distance function for time-series clustering, the concept of incorporating information about error in the distance function can be used in developing a distance function in any clustering application in which data that is clustered in other ways.

The problem

We describe the problem being solved in the following new way:

Let m objects (e.g., seasonalities) be represented by m data points (e.g., sales demand rates) in n-dimensional space (e.g., representing n weeks) . Each of these m data points represents a Gaussian distribution of data values, which defines the average value of each data point and also specifies the standard error associated with each data point.

We group these data points into clusters so that it is likely that data points (sales demand rates) in the same cluster are similar to (close to) each other, and data points in different clusters are unlikely to be close to each other, the likelihood being defined with respect to the Gaussian distributions represented by the data points. In contrast to other clustering techniques, two data points that differ significantly in their means may belong to the same cluster if they have high errors associated with them, and two data points that do not differ much in their means might belong to different clusters if they are well-calculated and have small errors.

Seasonality forecasting

As one example, we consider the forecasting of seasonality for retailers based on sales data from previous years. Seasonality is defined as the underlying demand for a group of similar merchandise items as a function of time of the year that is independent of external factors like changes in price, inventory,

and promotions. Seasonality for an item is expected to behave consistently from year to year.

In the retail industry, it is important to understand the seasonal behaviors in the sales of items to correctly forecast demand and make appropriate business decisions with respect to each item. Different items may be characterized by different seasonalities for a year. Yet, to reduce random variability and make better demand forecasts, it is desirable to group items with similar seasonal behavior together.

In this application, the objects to be clustered are the seasonalities of different groups of retail items. Effective clustering of retail items yields better forecasting of seasonal behavior for each of the retail items. Therefore, we identify a set of clusters of seasonalities that model most of the items sold by the retailer and relate each item to average seasonality of the cluster.

We model seasonality forecasting as a time-series clustering problem in the presence of errors, and we discuss experimental results on retail industry data. We have discovered meaningful clusters of seasonality that have not been uncovered by clustering methods that do not use the error information.

We assume that external factors, like price, promotions, and inventory, have been filtered from our data, and that the data has been normalized so as to compare sales of different items on the same scale. After normalization and filtering for external factors, the remaining demand rate of an item is determined by its Product Life Cycle (PLC) and seasonality. PLC is defined as the base

demand of an item over time in the absence of seasonality and all other external factors.

Filtering product life cycle effects

As shown in figure 1, for example, in a typical PLC 14, an item is introduced on a certain date 10 and removed from stores on a certain date 12. PLC 14 is a curve between the introduction date and removal date. The shape of the curve is determined by the duration of time an item is sold 16 and also the nature of the item. For example, a fashion item (right-hand curve 15) will sell out faster a non-fashion item (left-hand curve 14). For simplicity, we assume the PLC value to be zero during weeks when the item is not sold.

Because sales of an item is a product of its PLC and its seasonality, it is not possible to determine seasonality just by looking at the sales data of an item. The fact that items having the same seasonality might have different PLCs complicates the problem.

For example, if both the items for which PLC curves are shown in figure 1 follow the same seasonality 20 as shown in figure 2, then the sales of the two items (non-fashion and fashion) will be as shown in the two different curves 22 and 24 shown in the right-hand side of figure 2. Curves 22 and 24 are respectively the products of curves 14, 15 of figure 1 and the seasonality curve 20 in figure 2.

The first step is to remove as much as possible of the PLC factor from the sales data so that only the seasonality factor remains.

Initially, based on prior knowledge from merchants, we group items that are believed to follow similar seasonality over an entire year. As shown in figure 3, the items in the set 30 follow similar seasonality but may be introduced and removed at different points of time during the year. The set typically includes items having a variety of PLCs that differ in their shapes and durations of time. The weekly average of all PLCs in this set is a somewhat flat curve 32 as shown on the right-hand side of figure 3, implying that the weekly average of PLCs for all items in the set can be assumed to be constant.

This implies that averaging of weekly filtered demand rates of all items in the set would nullify the effect of PLCs and would correspond to the common value of seasonality indices for the items in the set to within a constant factor. The constant factor can be removed by appropriate scaling of weekly sales averages. Although the average values obtained above give a reasonably good estimate of seasonality, they will have errors associated with them depending on how many items were used to estimate seasonality and also on how well spread are their PLCs.

Error information

Let σ_j^2 be the variance of all the filtered demand rates for a set (e.g., one of the sets described above) of retail items in week j. To be clear, we are capturing the variance of the data values that represent the units of sales of the various items for a given week j as found in, say, one year of historical sales data. If there are a total of m items in this set, then an estimate of s_j, which is the standard

error in the estimation of the seasonality index for week j, is given by the following equation.

$$S_j = \frac{\sigma_j}{\sqrt{m}} \quad (1)$$

By seasonality index for week j, we mean the factor that represents the seasonality effect for all m items in the set relative to the baseline demand value. So, if the baseline demand is 20 units in week j and the seasonality index in week j is 1.3 then we expect the sales data in week j to be 26 units.

The above procedure provides a large number of seasonal indices, one for each set of retail items of merchandise, along with estimates of associated errors. We group these seasonal indices into a few clusters based on the average values and errors as calculated above, thus associating each item with one of these clusters of seasonal indices. The seasonality index of a cluster is defined as a weighted average of seasonalities present in the cluster as shown by equation 3 below. The cluster seasonality obtained in this way is much more robust because of the large number of seasonalities used to estimate it.

To summarize the discussion to this point, based on prior knowledge about which items, of the thousands of items that appear in a merchant's hierarchy of items, follow similar seasonality, we group items into sets of items that have similar seasonality. We estimate the seasonality for each of these sets and also estimate the error associated with our estimate of seasonality. This provides us with a large number of seasonalities along with

associated errors. We cluster these seasonalities into a few clusters, then calculate the average seasonality of each cluster. The final seasonality of an item is the seasonality of the cluster to which it belongs.

Generating seasonality forecasts by clustering without incorporating the errors information would disregard the fact that we do not know how much confidence we have in each seasonality. If we have high confidence in the estimate of a seasonality, then we would like to be more careful in assigning it to a cluster. On the other hand, if we have little confidence in the estimate of a seasonality, then its membership in a cluster does not have high significance in the cumulative average seasonality of the cluster. Errors or the associated probability distributions capture the confidence level of each seasonality and can be used intelligently to discover better clusters in the case of stochastic data.

Representation of time-series

Time-series are widely used in representing data in business, medical, engineering, and social sciences databases. Time-series data differs from other data representations in the sense that data points in a time-series are represented by a sequence typically measured at equal time intervals.

Stochastic data can be represented by a time-series in a way that models errors associated with data.

A time-series of stochastic data sampled at k weeks is represented by a sequence of k values. In our application, where we assume, for example, that the k samples are independent of each other and are each distributed according to one-dimensional Gaussian distribution, we represent a time-series A as:

$A = \{(\mu_1, s_1, w_1), (\mu_2, s_2, w_2), ..., (\mu_k, s_k, w_k)\}$ where the stochastic data of the i^{th} sample of A is normally distributed with mean μ_i and standard deviation s_i. w_i is a weight function to give relative importance to each sample of the time-series. Weights are chosen such that $\sum_i (\mu_i * w_i) = k$ so as to express time-series on the same scale. This normalization is important in the following sense. First of all, these weights reflect the relative, not absolute, importance of each sampled value. Secondly, the normalization converts the data into a unit scale thereby comparing differences in the shapes of time-series and not in the actual values. In figure 4, for example, the normalization will facilitate putting the time-series A and the time-series B in the same cluster and the time-series C in a separate cluster.

Although one can experiment with different weights for different sample values of a time-series, for simplicity, we assume that all the k sample values of a time-series have equal weight. Henceforth, we will work with the following compact representation of time-series. $A = \{(\mu'_1, s'_1), (\mu'_2, s'_2), ..., (\mu'_k, s'_k)\}$, where $\mu'_i = \mu_i * w$ and $s'_i = s_i * w$, for $i = 1,2,...,k$ where $w = \frac{k}{\sum_i \mu_i}$.

In the case of seasonality forecasting, we may have k equal to 52 corresponding to 52 weeks in a year, and μ_i be the estimate of the seasonality index for i^{th} week. s_i represents the standard error in the estimated value of μ_i.

We have assumed that sales data that are used to estimate means and errors of the respective k sample values of a time-series come from independent distributions. Although we might observe some level of independence, complete independence is not possible in real data. Especially in time-series data, one expects a positive correlation in consecutive sample values. Therefore, while incorporating the concept of dependence can be difficult, it can improve the test statistic for distance and subsequently give a more accurate measure of the distance function. In the case of seasonality forecasting, we deal with seasonality values that are obtained by taking the average of sales data of items having different PLCs. These PLCs are pretty much random for a large sample of data and therefore averaging over these random PLCs may dampen the effect of dependency among different samples. This implies that dependency is not a serious issue in our application, a proposition that is also observed experimentally.

Distance function

As in other clustering methods, we make a basic assumption that the relationship among pairs of seasonalities in a set of n seasonalities is described by an nXn matrix containing a measure of dissimilarity between the i^{th} and the j^{th} seasonalities. In clustering parlance, the measure of dissimilarity is referred to as a

distance function between the pair of seasonalities. Various distance functions have been considered for the case of deterministic data. We have developed a probability-based distance function in the case of multidimensional stochastic data.

Consider two estimated seasonality indices $A_i = \{(\mu_{i1}, s_{i1}), (\mu_{i2}, s_{i2}), \ldots, (\mu_{ik}, s_{ik})\}$ and $A_j = \{(\mu_{j1}, s_{j1}), (\mu_{j2}, s_{j2}), \ldots, (\mu_{jk}, s_{jk})\}$. A_i and A_j are the estimated time-series of two seasonalities based on historically observed sales for corresponding sets of items. Let the corresponding true seasonalities be $\{\bar{\mu}_{i1}, \bar{\mu}_{i2}, \ldots, \bar{\mu}_{ik}\}$ and $\{\bar{\mu}_{j1}, \bar{\mu}_{j2}, \ldots, \bar{\mu}_{jk}$. This means that the μ's are the observed means that are associated with the true means of $\bar{\mu}$'s.

We define similarity between two seasonalities (time-series) as the probability that the two seasonalities might be the same. Two seasonalities are considered the same if the corresponding μ's are close with high significance with respect to the associated errors. In other words, if we define the null hypothesis H_o as $A_i = A_j$ then similarity between A_i and A_j is the significance level of this hypothesis. Here, $A_i = A_j$ means corresponding $\bar{\mu}_{il} = \bar{\mu}_{jl}$ for $l = 1, \ldots, k$. The distance or dissimilarity d_{ij} between A_i and A_j is defined as (1 - similarity). In other words, d_{ij} is the probability of rejecting the above hypothesis. This distance function satisfies the following desirable properties.

$dist(A,B) = dist(B,A)$

$dist(A,B) \geq 0$

$dist(A,A) = 0$

$dist(A,B) = 0 \Rightarrow A = B$

The statistic for the test of the above hypothesis is $\sum_{l=1}^{k}\left(\frac{\mu_{il}-\mu_{jl}}{s_l}\right)^2$

where s_l is the pooled variance for week l defined as [redacted]

Under a Gaussian assumption, the above statistic follows a Chi-Square distribution with k-1 degrees of freedom. Therefore, the distance d_{ij} which is the significance level of rejecting the above hypothesis is given by the following equation.

$$d_{ij} = ChiSqr\_PDF\left(\sum_{l=1}^{k}\left(\frac{\mu_{il}-\mu_{jl}}{s_l}\right)^2, k-1\right) \qquad (2)$$

We use this distance between pairs of seasonality indices as the basis for clustering them by putting pairs that have low distances in the same cluster and pairs that have high distances in different clusters. We use a confidence interval (e.g., 90%) to find a threshold distance. If the distance between two seasonalities is less than the threshold distance value, we put them in the same cluster.

Merging time-series

Here we define a 'merge' operation to combine information from a set of time-series and produce a new time-series that is a compromise between all the time-series used to produce it. In seasonality forecasting, the time-series are sets of seasonalities and the new time-series represents the weighted average seasonality for a cluster of seasonalities. The shape of the resulting time-series depends not only on the sample values of individual time-series but also on errors associated with individual time-series.

Given r time-series $A_i = \{(\mu_{i1}, s_{i1}), (\mu_{i2}, s_{i2}), \ldots, (\mu_{ik}, s_{ik})\}, i = 1,2,\ldots r$ then the resulting time-series $C = \{(\mu_1, s_1), (\mu_2, s_2), \ldots, (\mu_k, s_k)\}$ is given by

$$\mu_j = \frac{\sum_{i=1}^{r} \frac{\mu_{ij}}{s_{ij}^2}}{\sum_{i=1}^{r} \frac{1}{s_{ij}^2}} \quad j = 1,2,\ldots,k \qquad (3)$$

$$s_j = \frac{1}{\sqrt{\sum_{i=1}^{r} \frac{1}{s_{ij}^2}}} \quad j = 1,2,\ldots,k \qquad (4)$$

As shown in the example of figure 5, consider two time-series 40 and 42 in which the curves 44, 46 represent the average values at each of 26 samples of time represented along the horizontal axis. The vertical line 48 for each of the samples represents the error associated with the sample at that time. The time-series 50 is the resulting sequence when the time-series 40 and 42 are merged in the manner discussed above. As can be seen, portions of each of the time-series 40 and 42 for which the errors are relatively smaller than in the other time-series serve more prominently in the merged time series 50. Therefore, if the series 50 represented a merged seasonality of two merchandise items as part of a cluster of items, a retailer could use the series 50 as a more accurate prediction of seasonality of the items that make up the cluster than would have been represented by clustered seasonalities that were merged without the benefit of the error information. This helps the retailer make better demand forecast as shown by experimental results.

Experimental Results

We generated data sets as described below. First we generated ten different kinds of PLCs from a Weibull distribution with different parameters. These PLCs differ in their peaks and shapes depending on the parameters used in the distribution, as shown in figure 6. The PLC data is randomly generated by choosing one of these 10 PLCs with equal probability and a uniformly distributed starting time over a period of one year.

Then we considered three different seasonalities corresponding to Christmas seasonality, summer seasonality, and winter seasonality respectively, as shown in figure 7. We generated sales data by multiplying the randomly generated PLC data with one of the three seasonalities. When we considered twelve instances, each instance produced by generating 25-35 PLCs, multiplying them by one of the above seasonalities and averaging weekly sales to obtain an estimate of corresponding seasonality, we obtained the values of seasonalities with associated estimates of errors as shown in figure 8.

As shown in figure 8, some of the seasonalities do not correspond to any of the original seasonalities and each has large errors. We ran the clustering method as described above and we got three clusters with cluster centers as shown in figure 9, where cluster centers are obtained by averaging of all PLCs in the same cluster according to equations 3 and 4. The resulting seasonalities match the original seasonalities of figure 7 well as shown in figure 9. We compared our result with standard hierarchical clustering that did

not consider the information about errors. The number of misclassifications were higher when we used hierarchical clustering with standard Euclidean distance without accounting for errors; as shown in figure 10.

Actual Data Results

Figure 11 shows an example based on actual retail sales data. Each of the seven time-series in this figure represents a seasonality that is obtained from a group of items that are known to follow similar seasonality. The clustering of the seven sets provides five clusters as shown in the figure.

The expressing of data in a way that incorporates error information, the expressing of a distance function based on the error information, the clustering of data sets using the distance function, the merging of clusters of data sets, the specific applications of the techniques to time-series data, including seasonality time-series for retail items, and other techniques described above, can be done using computers or machines controlled by software instructions.

For example, as shown in figure 12, the software would be stored in memory or on a medium or made available electronically, and would enable the computer or machine to perform the following sequence, for example, in the context of a retailer making pricing or inventory decisions with respect to retail items. After filtering and normalizing the historical time-sequence sales data for a number of items, in step 80, the data would be grouped and processed in step 82 to reduce the effect of the product life cycle

factor. This would include grouping items known to have similar seasonality, scaling, and determining variance and standard errors with respect to each group. The seasonality indices for each of the groups would then be expressed in a representation (step 84) that captures both (a) the mean values of seasonality for a set of items for each period and (b) the statistical or error information that represents the confidence level with respect to the mean values of seasonality. The software would analyze the various seasonality time-series using the distance functions (step 86) to cluster the time-series, and then would use the clustered time-series as part of the basis for decision-making (step 88). Additional information about the collection of such information and the making of such decisions is found in United States patent applications 09/263,979, filed March 5, 1999, 09/826,378, filed April 4, 2001, and 09/900,706, filed July 6, 2001, all incorporated by reference here.

Other implementations are within the scope of the following claims.

CLAIMS

1. A method comprising

receiving a set of data containing values associated with respective data points, the values associated with each of the data points being characterized by a distribution,

expressing the values for each of the data points in a form that includes information about the distribution of the values for each of the data points, and

using the distribution information in clustering the set of data with at least one other set of data containing values associated with data points.

2. The method of claim 1 in which the respective data points are related in a time-sequence.

3. The method of claim 1 in which the data points relate to a seasonality of at least one item of commerce.

4. The method of claim 1 in which each of the sets of data relates to seasonalities of items of commerce.

5. The method of claim 4 in which the items of commerce comprise retail products, the data points relate to times during a season, and the values associated with each of the data points correspond to respective ones of the retail products.

6. The method of claim 5 also including determining statistical measures of the variability of the values with respect to the data point.

7. The method of claim 1 in which the data is expressed in a form that includes a mean of the values associated with a data point and a statistical measure of the distribution with respect to the mean.

8. The method of claim 7 in which the statistical measure comprises a standard deviation.

9. The method of claim 1 in which the clustering of data includes measuring a distance between pairs of the sets of data.

10. The method of claim 9 in which the distance is measured based on the means and variances at the data points.

11. The method of claim 10 in which the distance is measured based on $d_{ij} = ChiSqr\_PDF\left(\sum_{l=1}^{k}\left(\frac{\mu_{il}-\mu_{jl}}{s_l}\right)^2, k-1\right)$

12. The method of claim 1 in which the distribution of the values comprises a Gaussian distribution.

13. The method of claim 1 in which the clustering of data includes merging the data sets belonging to a cluster using a weighted average.

14. The method of claim 1 also including merging the seasonalities of the data sets belong to a cluster

15. A machine-accessible medium that when accessed results in a machine effecting actions comprising:

receiving a set of data containing values associated with respective data points, the values associated with each of the data points being characterized by a distribution,

expressing the values for each of the data points in a form that includes information about a distribution of the values for each of the data points, and

using the distribution information in clustering the set of data with at least one other set of data containing values associated with data points.

16. A method comprising

receiving sets of data, each of the sets containing values associated with respect data points, the values associated with each of the data points being characterized by a distribution,

evaluating a distance function that characterizes the similarity or dissimilarity of at least two of the sets of data, the distance function including a factor based on the distributions of the values in the sets, and

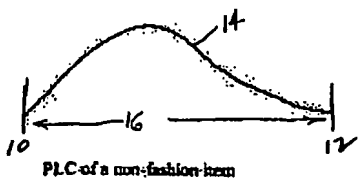
using the evaluation of the distance function as a basis for clustering of the sets of data.

17. A method comprising

receiving data that represents seasonality time-series for each of a set of retail items, the data also representing error

information associated with data values for each of a series of time points for each of the items, and

forming composite seasonality time-series based on respective clusters of the retail item seasonality time-series, the composites formed based in part on the error information.



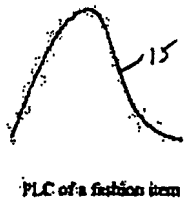


Figure 1

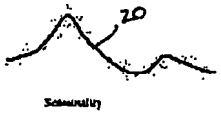


Figure 2

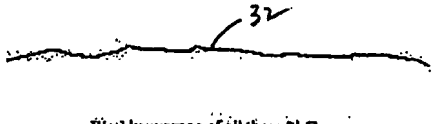


Figure 3

Figure 4

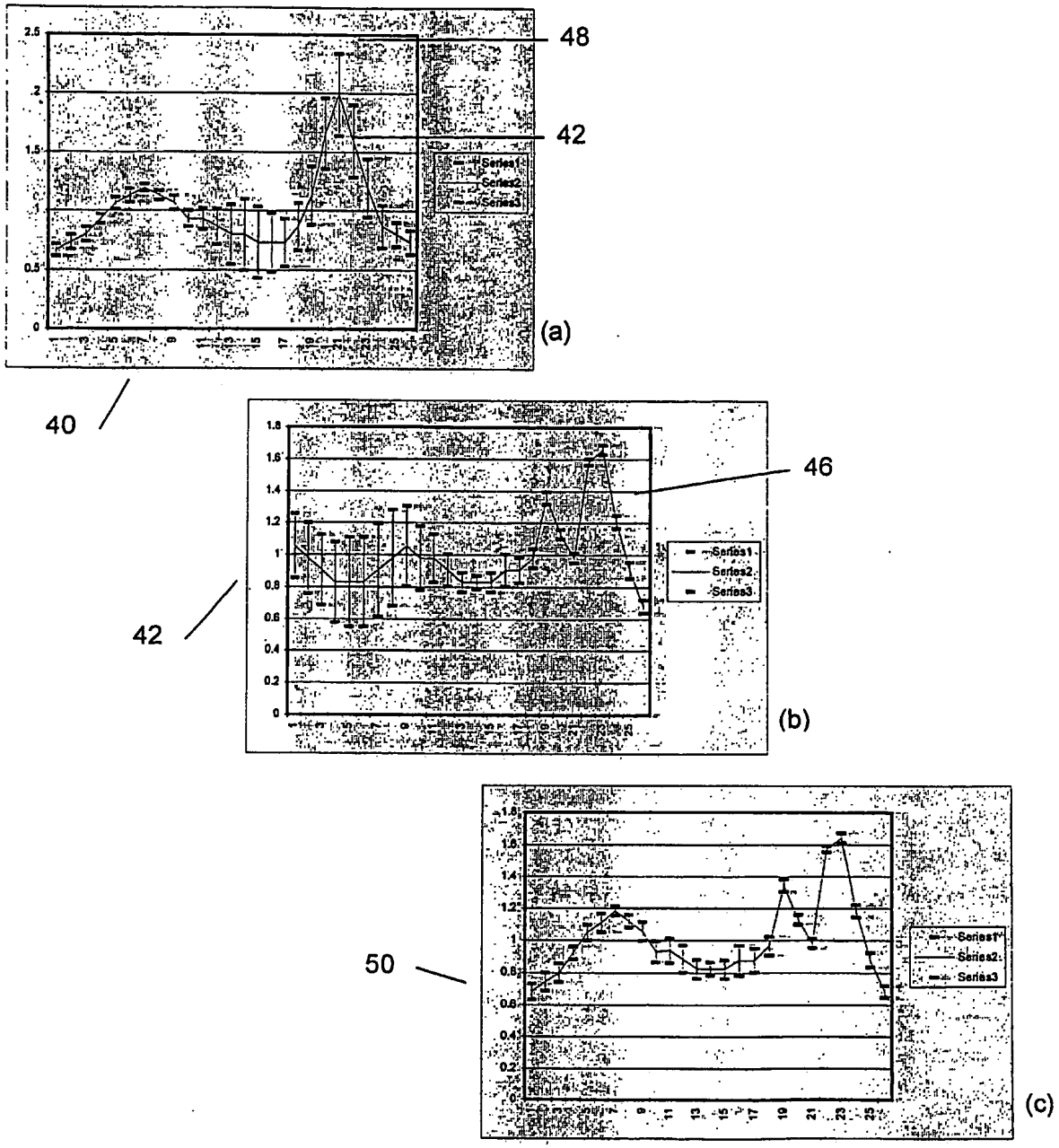


Figure 5

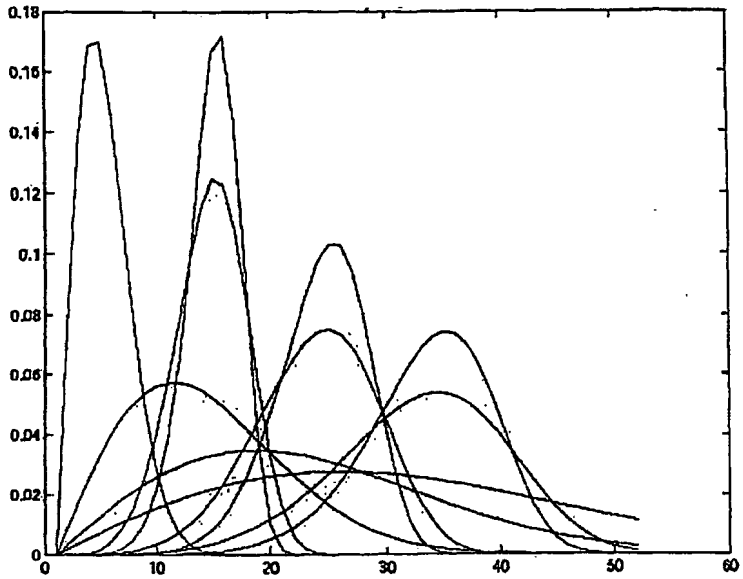


Figure 6

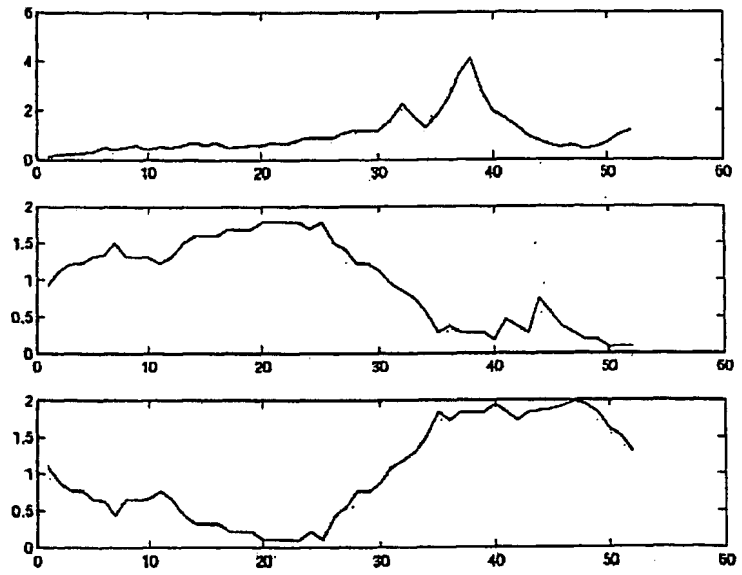


Figure 7

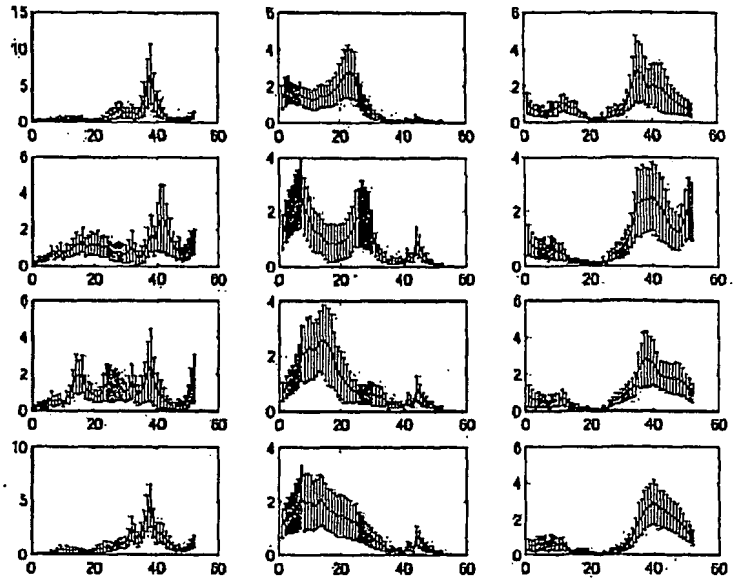


Figure 8

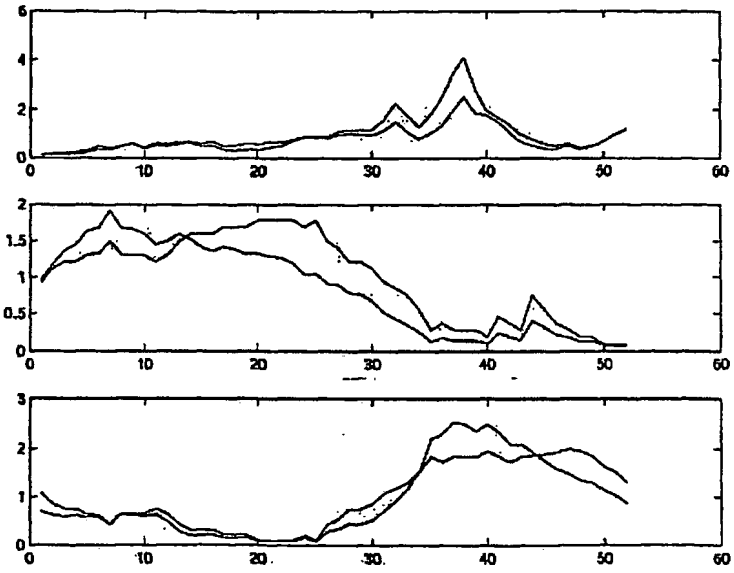


Figure 9

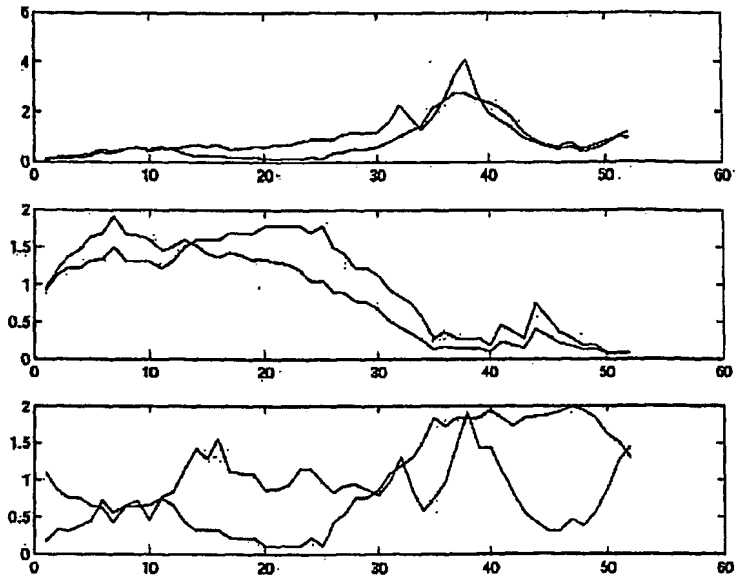

6
4
2
0
0 10 20 30 40 50 60
2
1.5
1
0.5
0
0 10 20 30 40 50 60
2
1.5
1
0.5
0
0 10 20 30 40 50 60

Figure 10

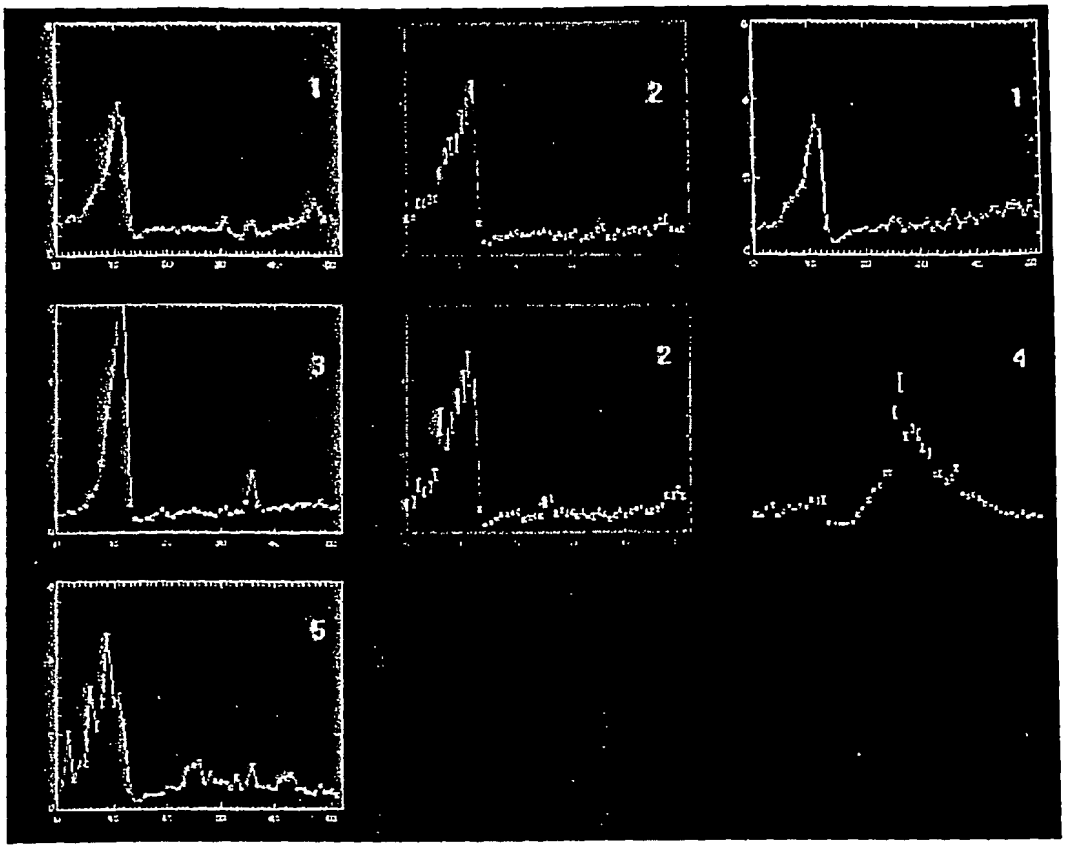


Figure 11

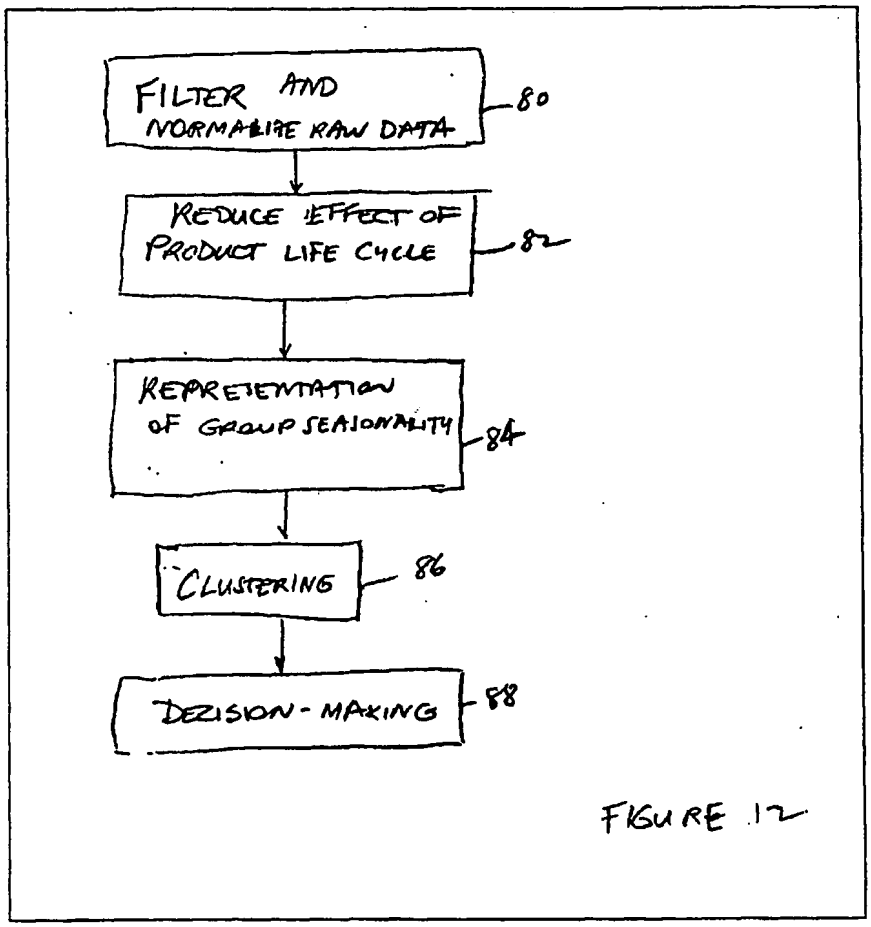

FILTER AND NORMALIZE RAW DATA
80
REDUCE EFFECT OF PRODUCT LIFE CYCLE
82
REPRESENTATION OF GROUP SEASONALITY
84
CLUSTERING
86
DECISION-MAKING
88

FIGURE 12